(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 11134334 A

(43) Date of publication of application: 21.05.99

(51) Int. Cl G06F 17/27
G06F 17/22
G06F 17/28
G06F 17/30

(21) Application number: 09296768

(22) Date of filing: 29.10.97

(71) Applicant: FUJITSU LTD

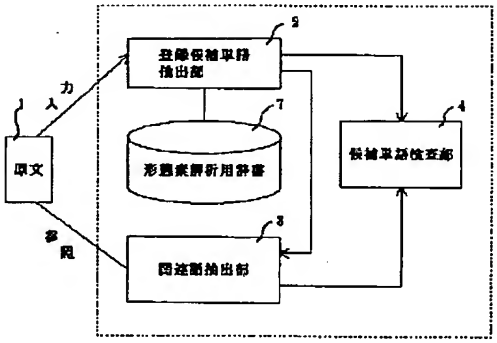
(72) Inventor: SATSUSANO YUKARI

(54) WORD REGISTERING DEVICE AND RECORDING MEDIUM

(57) Abstract:

PROBLEM TO BE SOLVED: To reduce a labor at the time of a registering work by extracting a word to be registered including an unregistered composite word.

SOLUTION: This device is provided with a morpheme analysis dictionary 7 for processing natural language, and an registration candidate word extracting part 2 for operating the morpheme analysis of a natural language sentence, extracts a composite word in which words which are not registered in the morpheme analysis dictionary 7 and nouns which are not registered in the dictionary 7 are continued, and judges the composite word whose frequency is high as a registration candidate word to be registered.



(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平11－134334

(43)公開日　平成11年(1999)５月21日

(51)Int.Cl.[6]	識別記号	FI		
G06F	17/27	G06F	15/38	E
	17/22		15/20	522L
	17/28		15/38	U
	17/30		15/403	330C

審査請求　未請求　請求項の数４　ＯＬ　（全 19 頁）

(21)出願番号　特願平9－296768

(22)出願日　平成９年(1997)10月29日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

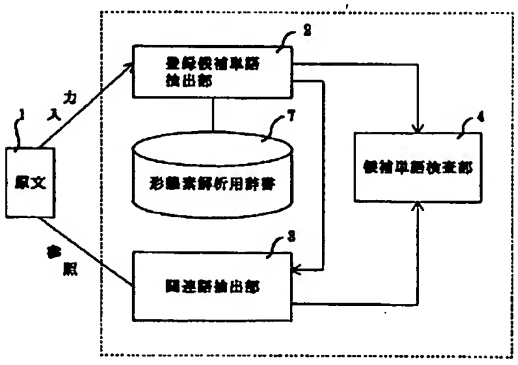
(72)発明者　颯々野　由香梨
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　弁理士　山谷　晧榮　（外２名）

(54)【発明の名称】　単語登録装置及び記録媒体

(57)【要約】
【課題】未登録複合語を含めた登録すべき単語を抽出し、登録作業時の労力を軽減すること。
【解決手段】自然言語を処理するための形態素解析用辞書７と、自然言語文を形態素解析し、前記形態素解析用辞書７に登録されていない単語及び該辞書７に登録されていない名詞類の連続した複合語を抽出して、頻度の高いものを登録すべき登録候補単語と判定する登録候補単語抽出部２とを備える。

本発明の原理説明図



1

【特許請求の範囲】

【請求項１】自然言語を処理するための形態素解析用辞書と、

自然言語文を形態素解析し、前記形態素解析用辞書に登録されていない単語及び該辞書に登録されていない名詞類の連続した複合語を抽出して、頻度の高いものを登録すべき登録候補単語と判定する登録候補単語抽出部とを備えることを特徴とした単語登録装置。

【請求項２】前記判定した登録候補単語を含む原文を検索し、前記形態素解析用辞書に登録されていない単語及び該辞書に登録されていない名詞類の連続した複合語を抽出する関連語抽出部を備えることを特徴とした請求項１記載の単語登録装置。

【請求項３】前記判定した登録候補単語を含む原文に対して、前記登録候補単語を取り入れる前の形態素解析結果と前記登録候補単語を取り入れた場合の形態素解析結果を比較して、解析誤りが起こっているかどうかを判定する候補単語検査部を備えることを特徴とした請求項１記載の単語登録装置。

【請求項４】コンピュータに、

自然言語文を形態素解析する解析手順と、

前記形態素解析結果から形態素解析用辞書に登録されていない単語を抽出する抽出手順と、

前記形態素解析結果から形態素解析用辞書に登録されていない名詞類の連続した複合語を抽出する抽出手順と、

前記抽出手順で抽出した単語及び複合語より頻度の高い語を登録候補単語と判定する判定手順と、

を実行するためのプログラムを格納したコンピュータ読取可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、自然言語を処理するための単語辞書に単語を追加登録する単語登録装置及び記録媒体に関する。

【０００２】

【従来の技術】日本語を形態素に分割する形態素解析は、自然言語処理の最も基本となる処理である。従来、形態素解析は、自然言語処理の様々なアプリケーションに用いられており、例えば、情報検索や文書中の誤りを発見する文書校正支援に用いられている。

【０００３】形態素解析にあたって、それに用いられる形態素解析用辞書は、形態素解析の性能を左右する重要な基本データである。この辞書中に単語が登録されていないと、解析が失敗したり、他の語として誤って解析されてしまう。例えば、事故や事件が起こった場合、関連記事を検索するための新しい単語を入力して検索するというニーズが増大しているが、関連の単語が辞書に入っていない場合、目的とする記事が検索できないという事態が生じる可能性がある。そのため、日々増加している新しい事象を表す単語を収集して、形態素解析用辞書に

2

追加することが重要である。しかし、新しい事象を表す単語は日々増加しているため、登録すべき単語を収集したり、テストする作業には多くの労力がかかっていた。

【０００４】従来、形態素解析用辞書に未登録語を登録する場合、形態素解析手段により入力文の解析を行い、その情報を基に入力文中の未登録語を知らせてユーザに登録を促すことが、特開平３－２４６６７３号公報に記載されていた。また、未登録語の出現回数を計算して、使用頻度の多いものから優先的に登録することが、特開昭６３－２０８１６７号公報に記載されていた。また、既知語の意味カテゴリを用いて未知語の意味カテゴリを推定して登録することが、特開平８－１６５９７号公報に記載されていた。また、関連情報辞書登録手段により、格の違いによる二重登録を排除し辞書量を少なくすることが、特開平６－１１９３７４号公報に記載されていた。

【０００５】

【発明が解決しようとする課題】前記のような従来のものは、次のような課題があった。

①：二つ以上の名詞類が連続している未登録複合語を抽出できるものではなかった。

【０００６】②：登録候補単語の関連である入力文中に含まれる頻度の低い未登録語を登録できるものではなかった。

③：登録すべき単語のテストを事前に行えるものではなかった。

【０００７】本発明は、このような従来の課題を解決し、未登録複合語を含めた登録すべき単語の抽出をし、登録作業時の労力を軽減し、更に登録すべき単語のテストを事前に行い、質のよい単語を半自動的に収集すること、また、登録すべき単語候補として選ばれた単語と関連のある語も同時に収集できるようにすることを目的とする。

【０００８】

【課題を解決するための手段】図１は本発明の原理説明図である。図１中、１は原文、２は登録候補単語抽出部、３は関連語抽出部、４は候補単語検査部、７は形態素解析用辞書である。

【０００９】本発明は前記従来の課題を解決するため次のように構成した。

（１）：自然言語を処理するための形態素解析用辞書７と、自然言語文を形態素解析し、前記形態素解析用辞書７に登録されていない単語及び該辞書７に登録されていない名詞類の連続した複合語を抽出して、頻度の高いものを登録すべき登録候補単語と判定する登録候補単語抽出部２とを備える。

【００１０】（２）：前記（１）の単語登録装置において、前記判定した登録候補単語を含む原文１を検索し、前記形態素解析用辞書７に登録されていない単語及び該辞書７に登録されていない名詞類の連続した複合語を抽

出する関連語抽出部３を備える。

【００１１】（３）：前記（１）の単語登録装置において、前記判定した登録候補単語を含む原文１に対して、前記登録候補単語を取り入れる前の形態素解析結果と前記登録候補単語を取り入れた場合の形態素解析結果を比較して、解析誤りが起こっているかどうかを判定する候補単語検査部４を備える。

【００１２】（４）：コンピュータに、自然言語文を形態素解析する解析手順と、前記形態素解析結果から形態素解析用辞書７に登録されていない単語を抽出する抽出手順と、前記形態素解析結果から形態素解析用辞書７に登録されていない名詞類の連続した複合語を抽出する抽出手順と、前記抽出手順で抽出した単語及び複合語より頻度の高い語を登録候補単語と判定する判定手順と、を実行するためのプログラムを格納したコンピュータ読取可能な記録媒体とする。

【００１３】（作用）前記構成に基づく作用を説明する。登録候補単語抽出部２で、自然言語文を形態素解析し、形態素解析用辞書７に登録されていない単語及び該辞書７に登録されていない名詞類の連続した複合語を抽出して、頻度の高いものを登録すべき登録候補単語と判定する。このため、頻度の高い未登録語だけでなく頻度の高い未登録複合語も登録候補単語として判定することができ、登録すべき語の抽出及び選択作業を軽減することができる。

【００１４】また、関連語抽出部３で、前記判定した登録候補単語を含む原文１を検索し、形態素解析用辞書７に登録されていない単語及び該辞書７に登録されていない名詞類の連続した複合語を抽出する。このため、頻度が低い語も関連語として原文から抽出し、その語も登録候補単語として取り入れることができる。

【００１５】さらに、候補単語検査部４で、前記判定した登録候補単語を含む原文１に対して、前記登録候補単語を取り入れる前の形態素解析結果と前記登録候補単語を取り入れた場合の形態素解析結果を比較して、解析誤りが起こっているかどうかを判定する。このため、登録する前にテストが行え、質のよい単語を収集することができる。

【００１６】また、自然言語文を形態素解析する解析手順と、前記形態素解析結果から形態素解析用辞書７に登録されていない単語を抽出する抽出手順と、前記形態素解析結果から形態素解析用辞書７に登録されていない名詞類の連続した複合語を抽出する抽出手順と、前記抽出手順で抽出した単語及び複合語より頻度の高い語を登録候補単語と判定する判定手順と、を実行するためのプログラムを格納したコンピュータ読取可能な記録媒体とする。このため、この記録媒体のプログラムをコンピュータにインストールすることで、頻度の高い未登録語だけでなく頻度の高い未登録複合語も登録候補単語として判定することができる単語登録装置を容易に提供することができる。

【００１７】

【発明の実施の形態】本発明の単語登録装置では、日々更新されるニュース記事やＷｅｂページ（インターネットのホームページ）等の記事を形態素解析し、登録すべき単語候補を抽出し、その語が登録した場合の解析のテストを行う機構を設けることで、登録すべき単語の抽出や登録作業時の労力を軽減するものである。また、登録すべき単語候補として選ばれた単語と同時に登録すべき関連語も原文から抽出し、その語も登録単語候補として取り入れる機能を備えるものである。

【００１８】図２～図１６は本発明の実施の形態を示した図である。以下、図２～図１６に基づいて本発明の実施の形態を説明する。

（１）：装置構成の説明

図２は装置構成図である。図２において、原文データ１が入力される単語登録装置には、登録候補単語抽出部２、関連語抽出部３、候補単語検査部４、単語登録部５、形態素解析エンジン６、形態素解析用辞書７が設けてある。

【００１９】原文データ１は、入力手段（図示せず）により入力される日々更新されるニュース記事やＷｅｂページ等の記事である。登録候補単語抽出部２は、形態素解析結果から登録候補単語を抽出するものである。関連語抽出部３は、登録候補単語を元に関連語を抽出するものである。候補単語検査部４は、元の解析結果と登録候補単語を取り入れた場合の解析結果を比較して、解析誤りが起こっているかどうかを判定するものである。単語登録部５は、ユーザに登録候補単語や関連語の検査結果を表示し、形態素解析用辞書７に格納するものである。形態素解析エンジン６は、形態素解析を行う処理部である。形態素解析用辞書７は、形態素解析に使用するための単語を登録しておくものである。

【００２０】（２）：全体の処理手順の説明

図３は全体の処理手順の説明図である。以下、図３の処理Ｓ１～処理Ｓ４に従って説明する。

【００２１】Ｓ１：決められた時間にダウンロード等で自動で入力された新聞記事等の原文データ１を登録候補単語抽出部２で、形態素解析し、その結果から登録候補単語を抽出し、処理Ｓ２に移る。

【００２２】Ｓ２：関連語抽出部３で、登録候補単語として選ばれた単語を含む元記事中に含む単語（関連語）を登録候補単語として選択し、処理Ｓ３に移る。

Ｓ３：候補単語検査部４で、登録候補単語及び関連語を登録した場合の形態素解析結果をテストし、その結果をユーザに提示し、処理Ｓ４に移る。

【００２３】Ｓ４：ユーザが登録すべき単語として指示した場合、単語登録部５で形態素解析用辞書７に登録して、この処理を終了する。

（３）：登録候補単語抽出部の処理の説明

図４は登録候補単語抽出部の処理の説明図である。以下、図４の処理Ｓ１１～処理Ｓ１４に従って説明する。

【００２４】Ｓ１１：登録候補単語抽出部２は、原文データ１に対して、形態素解析エンジン６と形態素解析用辞書７を用いて形態素解析を行い、処理Ｓ１２に移る。

Ｓ１２：登録候補単語抽出部２は、形態素解析結果から未登録語を抽出して、未登録語頻度表を作成し、処理Ｓ１３に移る。

【００２５】Ｓ１３：登録候補単語抽出部２は、形態素解析結果から名詞類の連続を抽出して、未登録複合語頻度表を作成し、処理Ｓ１４に移る。

Ｓ１４：登録候補単語抽出部２は、それぞれ作成した頻度表の頻度の上位のものを登録候補単語リストに登録して、この処理を終了する。

【００２６】（４）：関連語抽出部の処理の説明

図５は関連語抽出部の処理の説明図である。以下、図５の処理Ｓ２１～処理Ｓ２３に従って説明する。

【００２７】Ｓ２１：関連語抽出部３は、登録候補単語を含む元の文の記事を検索し、処理Ｓ２２に移る。

Ｓ２２：関連語抽出部３は、その記事中に未登録語頻度表、未登録複合語頻度表に含まれる語が存在するかを判定し、処理Ｓ２３に移る。

【００２８】Ｓ２３：関連語抽出部３は、各頻度表に含まれる語があれば、それを関連語として抽出し、登録候補単語リストに追加して、この処理を終了する。

（５）：候補単語検査部の処理の説明

図６は候補単語検査部の処理の説明図である。以下、図６の処理Ｓ３１～処理Ｓ３４に従って説明する。

【００２９】Ｓ３１：候補単語検査部４は、登録候補単語リストから候補単語辞書を作成すし、処理Ｓ３２に移る。

Ｓ３２：候補単語検査部４は、登録候補単語を含む原文に対して、元の形態素解析用辞書と候補単語辞書を用いて、形態素解析をし、処理Ｓ３３に移る。

【００３０】Ｓ３３：候補単語検査部４は、元の形態素解析結果と登録候補単語を取り入れた場合の形態素解析結果を比較して、解析誤りが起こっているかどうかを判定し、処理Ｓ３４に移る。

【００３１】Ｓ３４：候補単語検査部４は、解析誤りが起こっている単語を登録候補単語リストから除外し、この処理を終了する。なお、解析誤りの例として、登録候補単語を取り入れた場合に他の部分（特に取り入れた登録候補単語の前後部分）が未登録語となる場合や逆に未登録語が増加する場合がある。

【００３２】（６）：単語登録部の処理の説明

図７は単語登録部の処理の説明図である。以下、図７の処理Ｓ４１～処理Ｓ４４に従って説明する。

【００３３】Ｓ４１：単語登録部５は、登録候補単語リストと元の形態素解析結果とそれに新たに登録した場合の形態素解析結果をユーザに提示し、処理Ｓ４２に移る。

Ｓ４２：ユーザが登録候補単語から登録すべき単語を選択し、処理Ｓ４３に移る。

【００３４】Ｓ４３：単語登録部５は、ユーザに単語の辞書上の登録情報を候補単語辞書から提示し、処理Ｓ４４に移る。

Ｓ４４：ユーザが候補単語辞書の内容をそのまま、あるいは修正して、単語登録部５で形態素解析用辞書７に登録し、この処理を終了する。

【００３５】（７）：具体例による説明

ａ：登録候補単語を登録する場合の説明

図８は登録候補単語を登録する場合の説明図（１）であり、図８（ａ）は一文の形態素解析例の説明、図８（ｂ）は未登録単語頻度表の説明である。図９は登録候補単語を登録する場合の説明図（２）であり、図９（ａ）は候補単語辞書の説明、図９（ｂ）は登録前の形態素解析結果の説明である。図１０は登録候補単語を登録する場合の説明図（３）であり、図１０（ａ）は「ヤンゴン」を登録した場合の形態素解析結果の説明、図１０（ｂ）はユーザが修正した候補単語辞書の説明である。

【００３６】以下は、いくつかの内容を含む新聞記事から登録単語を抽出する例を図８～図１０により説明する。まず、登録候補単語抽出部２において、原文を形態素解析する。形態素解析の結果は、例えば、図８（ａ）のように、文が形態素単位に分割され、それぞれの品詞、詳細品詞、表記が出力される。

【００３７】登録候補単語抽出部２では、形態素解析の解析結果から、詳細品詞が「未登録語」となっている単語を収集し、図８（ｂ）のように頻度が記入された未登録単語頻度表を作成する。

【００３８】登録候補単語抽出部２は、原文の数に応じて頻度が上位であるものを登録すべき候補の単語として抽出する。例えば、ここで頻度が「１０」で頻度の高い「ヤンゴン」を登録候補単語として抽出する。候補単語検査部４では、登録候補単語である「ヤンゴン」に仮の品詞として、普通名詞を付与し、候補単語辞書を作成する。この候補単語辞書は、図９（ａ）のように表記、品詞、詳細品詞が設けられている候補単語検査部４では、登録候補単語が出現している文を元の形態素解析用辞書７と登録候補単語を取り入れた辞書を使って解析し直して、その結果を出力する。例えば、登録候補単語「ヤンゴン」を含む文が次のものであったとする。

【００３９】「ミャンマーの首都ヤンゴンで学生のデモが始まった。」この文に対して、「ヤンゴン」を登録する前の形態素解析結果は、図９（ｂ）であり、「ヤンゴン」を登録した場合の形態素解析結果は、図１０（ａ）である。図９（ｂ）において、未登録語であった「ヤンゴン」は、図１０（ａ）においては普通名詞となり他の単語にも未登録語が含まれていない。このため「ヤンゴ

ン」を登録した場合の結果に解析誤りは含まれていない。

【００４０】候補単語検査部４は、この結果を単語登録部５に渡し、ユーザに提示する。ユーザは、この結果を確認し、「ヤンゴン」を辞書に登録することを指示する。ここで「ヤンゴン」は、地名であるので、ユーザは、詳細品詞を「地名」に修正する。即ち、図１０（ｂ）のように候補単語辞書の情報を修正して形態素解析用辞書７に登録する。

【００４１】ｂ：未登録複合語頻度表を作成する場合の説明

図１１は未登録複合語頻度表を作成する場合の説明図（１）であり、図１１（ａ）は未登録複合語頻度表の説明、図１１（ｂ）は候補単語辞書の説明、図１１（ｃ）は登録前の形態素解析結果の説明である。図１２は未登録複合語頻度表を作成する場合の説明図（２）であり、図１２（ａ）は登録した後の形態素解析結果の説明、図１２（ｂ）はユーザが修正した候補単語辞書の説明である。

【００４２】登録候補単語抽出部２で、形態素解析結果から未登録単語頻度表以外に、未登録複合語頻度表を作成するものである。これは、二つ以上の名詞類（名詞、接頭語、接尾語、「・」、「//」、「＝」、動詞の連用形等）が連続しているものを取り出し、その頻度を調査したものである。

【００４３】ここで、未登録複合語頻度表が、図１１（ａ）のように得られたとする。なお、図１１（ａ）において、形態素の区切りは「／」で表している。ここでは、頻度が「１２」と高い「オーム／真理／教」を登録候補単語として抽出したとする。候補単語検査部４では、図１１（ｂ）のように「オーム真理教」に仮の品詞として、普通名詞を付与し、この「オーム真理教」が出現した文において形態素解析のテストを行う。

【００４４】候補単語検査部４では、登録候補単語が出現している文を元の形態素解析用辞書７と登録候補単語を取り入れた辞書を使って解析し直して、その結果を出力する。ここで、「オーム真理教」を含む原文が次のものであったとする。

【００４５】「オーム真理教の信者の林春男容疑者がきょう逮捕されました。」これを「オーム真理教」を一語として登録する前の形態素解析結果は、図１１（ｃ）に示してあり、登録した後の形態素解析結果は、図１２（ａ）に示してある。図１１（ｃ）と図１２（ａ）のように、「オーム真理教」を登録した場合の結果に解析誤りは含まれていないので、候補単語検査部４は、この結果を単語登録部５に渡し、ユーザに提示する。

【００４６】ユーザは、この結果を確認し、「オーム真理教」を辞書に登録することを指示する。ここで「オーム真理教」は、固有名詞であるので、ユーザは、詳細品詞を「固有名詞」に修正する。即ち、図１２（ｂ）のように候補単語辞書の情報を修正して形態素解析用辞書７に登録する。

【００４７】ｃ：関連語を登録する場合の説明

図１３は関連語を登録する場合の説明図（１）であり、図１３（ａ）は候補単語辞書（関連語）の説明、図１３（ｂ）は登録前の形態素解析結果の説明である。図１４は関連語を登録する場合の説明図（２）であり、図１４（ａ）は「國林長」を登録した場合の形態素解析結果の説明、図１４（ｂ）は「國林長官狙撃事件」を登録した場合の形態素解析結果の説明である。図１５は関連語を登録する場合の説明図（３）であり、図１５（ａ）はユーザが修正した候補単語辞書の説明、図１５（ｂ）は登録前の形態素解析結果の説明である。図１６は関連語を登録する場合の説明図（４）であり、図１６（ａ）は「アウン・タン・スー・チー」を登録した場合の形態素解析結果の説明、図１６（ｂ）はユーザが修正した候補単語辞書の説明である。

【００４８】前記具体例ａ、ｂのように「ヤンゴン」と「オーム真理教」を登録候補単語として抽出した場合、関連語抽出部３では、以下のように処理を行う。関連語抽出部３では、登録候補単語を含む記事中に含まれる頻度の低い未登録語や未登録複合語を選択する。これにより、以下の選択結果が得られたとする。

【００４９】「國林長」

「國林長／官／狙撃／事件」

「アウン／・／タン／・／スー／・／チー」

以上の関連語を登録候補単語リストに追加し、候補単語検査部４でテストを行う。候補単語検査部４では、以上の登録候補単語と関連語に仮の品詞として、普通名詞を付与し、それぞれの語が出現した文において形態素解析のテストを行う。例えば、関連語から図１３（ａ）のような候補単語辞書（関連語）を作る。

【００５０】候補単語検査部４では、登録候補単語が出現している文を元の形態素解析用辞書７と登録候補単語を取り入れた辞書を使って解析し、その結果を出力する。これは例えば、関連語を含む文が次のようであったとする。

【００５１】「警察庁の國林長官狙撃事件の捜査をめぐる対応が適切でない。」

「アウン・タン・スー・チーさんの勢力とは一線を画している。」

・「國林長官狙撃事件」を含む文の形態素解析結果は、登録前は図１３（ｂ）となり、「國林長」を登録した場合は図１４（ａ）となり、「國林長官狙撃事件」を登録した場合は図１４（ｂ）となる。

【００５２】ここで、「國林長」と「國林長官狙撃事件」を登録した場合は、いずれも解析誤りが起こっていないので、候補単語検査部４は、その結果を単語登録部５に渡し、ユーザに提示する。ユーザは、図１４（ａ）と図１４（ｂ）の形態素解析結果から、「國林長官狙撃

事件」を登録する方が正しいと判断し、「國林長官狙撃事件」を登録するとユーザが指示する。

【００５３】この場合、品詞は固有名詞なので、ユーザは、図１５（ａ）のように候補単語辞書の詳細品詞を「普通名詞」から「固有名詞」に修正し、単語登録部５で形態素解析用辞書７に取り込むようにする。

【００５４】・次に「アウン・タン・スー・チー」を登録する前と登録した後の形態素解析結果は、図１５（ｂ）と図１６（ａ）のようになる。ここで、「アウン・タン・スー・チー」を登録した場合は、解析誤りが起こっていないので、その結果を単語登録部５に渡し、ユーザに提示する。ユーザは、この結果を確認し、「アウン・タン・スー・チー」を形態素解析用辞書７に登録することを指示する。ここで、「アウン・タン・スー・チー」は人名であるので、ユーザは、候補単語辞書の詳細品詞を「普通名詞」から「人名」に修正し、単語登録部５で形態素解析用辞書７に取り込むようにする。

【００５５】以上実施の形態で説明したように、登録すべき単語の抽出および選択が軽減され、更に登録すべき単語のテストを事前に行えるので、質の良い単語を半自動的に収集できる。また、関連のある語も同時に収集することが可能となる。

【００５６】（８）：プログラムのインストールの説明登録候補単語抽出部２、関連語抽出部３、候補単語検査部４、単語登録部５、形態素解析エンジン６は実際にはプログラムで構成でき、主制御部（ＣＰＵ）が実行するものであり、主記憶に格納されているものである。これらのプログラムは、一般的な、パーソナルコンピュータ、ワークステーション等のデータ処理装置（コンピュータ）で処理されるものである。これらのコンピュータは、主制御部、主記憶、ハードディスク等のファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。

【００５７】このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピー、光磁気ディスク等の可搬型の記録媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記憶媒体に対して、アクセスするためのドライブ装置を介して、或いは、ＬＡＮ等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。

【００５８】

【発明の効果】以上説明したように、本発明によれば次のような効果がある。

（１）登録候補単語抽出部で、自然言語文を形態素解析し、形態素解析用辞書に登録されていない単語及び該辞書に登録されていない名詞類の連続した複合語を抽出して、頻度の高いものを登録すべき登録候補単語と判定するため、頻度の高い未登録語だけでなく頻度の高い未登録複合語も登録候補単語として判定することができ、登録すべき語の抽出及び選択作業を軽減することができる。

【００５９】（２）：関連語抽出部で、登録候補単語を含む原文を検索し、形態素解析用辞書に登録されていない単語及び該辞書に登録されていない名詞類の連続した複合語を抽出するため、頻度が低い単語及び複合語も関連語として原文から抽出し、その語も登録候補単語として取り入れることができる。

【００６０】（３）：候補単語検査部で、登録候補単語を含む原文に対して、前記登録候補単語を取り入れる前の形態素解析結果と前記登録候補単語を取り入れた場合の形態素解析結果を比較して、解析誤りが起こっているかどうかを判定するため、登録する前にテストが行え、質のよい単語を収集することができる。

【００６１】（４）：自然言語文を形態素解析する解析手順と、前記形態素解析結果から形態素解析用辞書に登録されていない単語を抽出する抽出手順と、前記形態素解析結果から形態素解析用辞書に登録されていない名詞類の連続した複合語を抽出する抽出手順と、前記抽出手順で抽出した単語及び複合語より頻度の高い語を登録候補単語と判定する判定手順と、を実行するためのプログラムを格納したコンピュータ読取可能な記録媒体とするため、この記録媒体のプログラムをコンピュータにインストールすることで、頻度の高い未登録語だけでなく頻度の高い未登録複合語も登録候補単語として判定することができる単語登録装置を容易に提供することができる。

【図面の簡単な説明】

【図１】本発明の原理説明図である。

【図２】実施の形態における装置構成図である。

【図３】実施の形態における全体の処理手順の説明図である。

【図４】実施の形態における登録候補単語抽出部の処理の説明図である。

【図５】実施の形態における関連語抽出部の処理の説明図である。

【図６】実施の形態における候補単語検査部の処理の説明図である。

【図７】実施の形態における単語登録部の処理の説明図である。

【図８】実施の形態における登録候補単語を登録する場合の説明図（１）である。

【図９】実施の形態における登録候補単語を登録する場合の説明図（２）である。

【図１０】実施の形態における登録候補単語を登録する場合の説明図（３）である。

【図１１】実施の形態における未登録複合語頻度表を作成する場合の説明図（１）である。

11　　12

【図１２】実施の形態における未登録複合語頻度表を作成する場合の説明図（２）である。

【図１３】実施の形態における関連語を登録する場合の説明図（１）である。

【図１４】実施の形態における関連語を登録する場合の説明図（２）である。

【図１５】実施の形態における関連語を登録する場合の説明図（３）である。

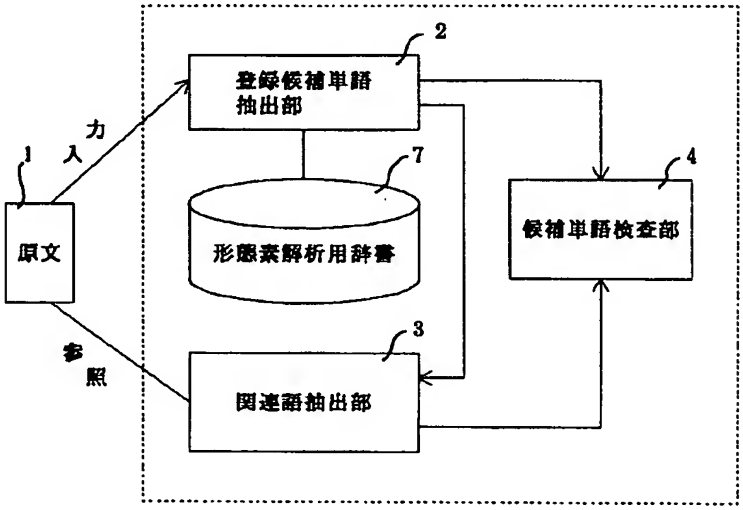
【図１６】実施の形態における関連語を登録する場合の説明図（４）である。

【符号の説明】

1　原文
2　登録候補単語抽出部
3　関連語抽出部
4　候補単語検査部
7　形態素解析用辞書

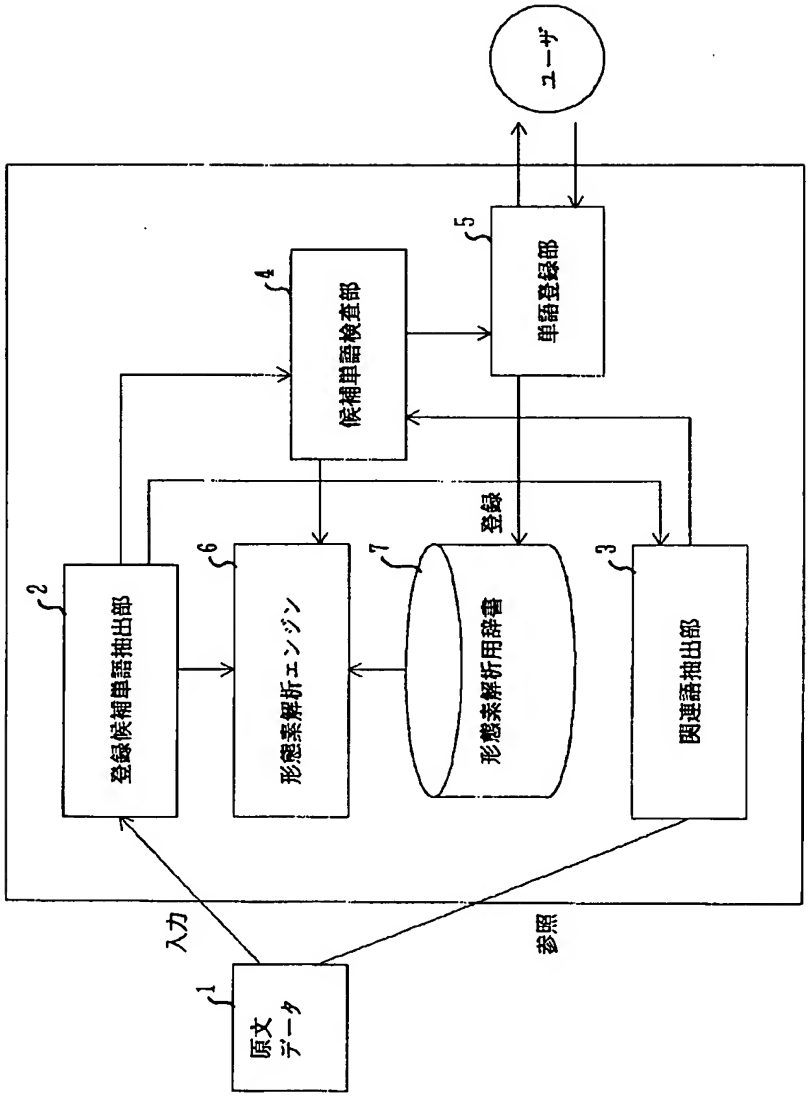
【図１】

本発明の原理説明図



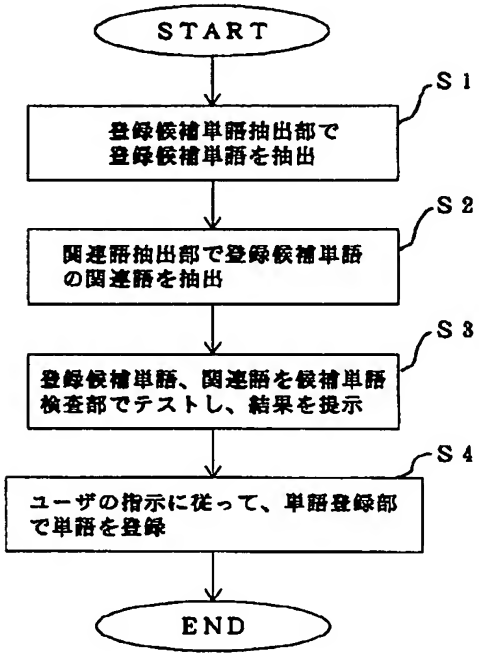
【図２】

装置構成図


ユーザ
単語登録部
候補単語検査部
登録
登録候補単語抽出部
形態素解析エンジン
形態素解析用辞書
関連語抽出部
入力
参照
原文データ
5
4
2
6
7
3
1

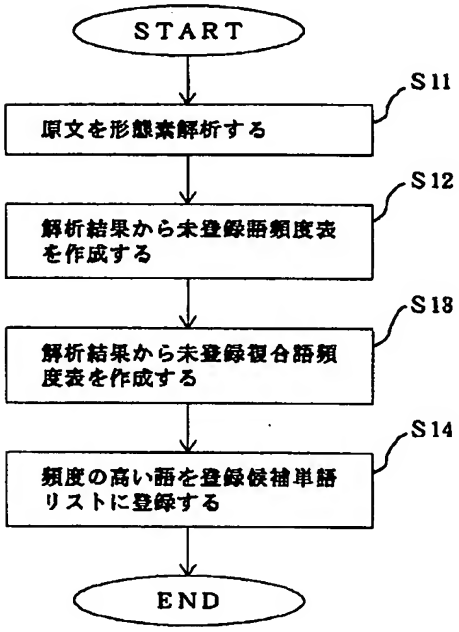
【図３】

全体の処理手順の説明図



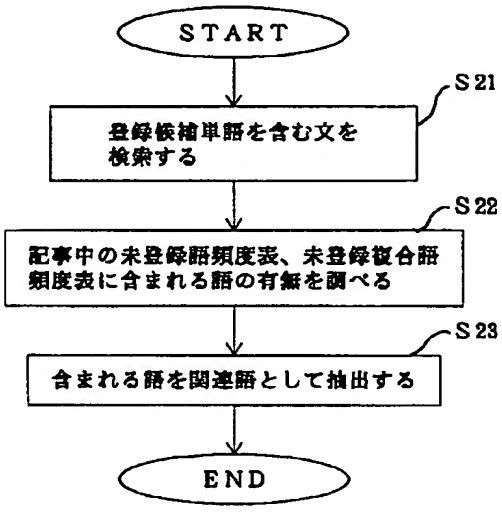
【図４】

登録候補単語抽出部の処理の説明図



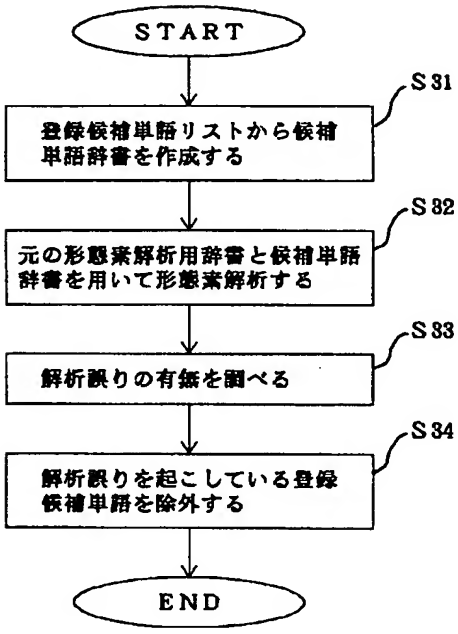
【図５】

関連語抽出部の処理の説明図



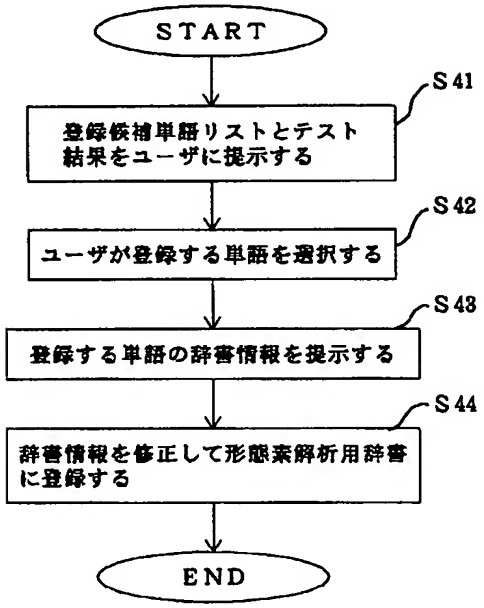
【図６】

候補単語検査部の処理の説明図



【図7】

単語登録部の処理の説明図



【図１３】

関連語を登録する場合の説明図（１）

（ａ）候補単語辞書（関連語）の説明

表記	品詞
國林長	普通名詞
國林長官狙撃事件	普通名詞
アウン・タン・スー・チー	普通名詞

（ｂ）登録前の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	固有名詞	警察庁
助詞	助詞「の」	の
名詞	未登録語	國林長
接尾語	接尾語	官
名詞	サ変名詞	狙撃
名詞	普通名詞	事件
助詞	助詞「の」	の
名詞	サ変名詞	捜査
助詞	格助詞	を
動詞	ラ行五段動詞	めぐ
動詞語尾	ラ行五段終止連	る
名詞	サ変名詞	対応
助詞	格助詞	が
形容動詞	ナ活用形容動詞	適切
形容動詞	形容動詞連用形	で
助動詞	ない	ない
形容詞語	形容詞終止連体	い
その他	句点	。

【図８】

登録候補単語を登録する場合の説明図（１）

（ａ）一文の形態素解析例の説明

品詞	詳細品詞	表記
名詞	サ変名詞	逮捕
動詞語尾	サ変未然形	さ
助動詞	れる	れ
助動詞	た終止連体形	た
助詞	助詞「の」	の
助詞	提題助詞	は
名詞	普通名詞	オーム
名詞	普通名詞	真理
名詞	普通名詞	教
助詞	助詞「の」	の
名詞	人称名詞	信者
助詞	助詞「の」	の
名詞	人名	林
名詞	人名	泰男
名詞	人称名詞	容疑者
助動詞	だ・です終止連	です
その他	句点	。

（ｂ）未登録単語頻度表の説明

頻度	品詞	詳細品詞	表記
１０	名詞	未登録語	ヤンゴン
５	名詞	未登録語	國林長
２	名詞	未登録語	アコヤ

【図９】

登録候補単語を登録する場合の説明図（２）

（ａ）候補単語辞書の説明

表記	品詞	詳細品詞
ヤンゴン	名詞	普通名詞

（ｂ）登録前の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	地名	ミャンマー
助詞	助詞「の」	の
名詞	普通名詞	首都
名詞	未登録語	ヤンゴン
助詞	格助詞	で
名詞	人称名詞	学生
助詞	助詞「の」	の
名詞	サ変名詞	デモ
助詞	格助詞	が
動詞	ラ行五段動詞	始ま
動詞語尾	ラ行五段連用形	っ
助動詞	た終止連体形	た
その他	句点	。

【図１０】

登録候補単語を登録する場合の説明図（３）

（ａ）「ヤンゴン」を登録した場合の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	地名	ミャンマー
助詞	助詞「の」	の
名詞	普通名詞	首都
名詞	普通名詞	ヤンゴン
助詞	格助詞	で
名詞	人称名詞	学生
助詞	助詞「の」	の
名詞	サ変名詞	デモ
助詞	格助詞	が
動詞	ラ行五段動詞	始ま
動詞語尾	ラ行五段連用形	っ
助動詞	た終止連体形	た
その他	句点	。

（ｂ）ユーザが修正した候補単語辞書の説明

表記	品詞	詳細品詞
ヤンゴン	名詞	地名

【図１１】

未登録複合語頻度表を作成する場合の説明図（１）

（ａ）未登録複合語頻度表の説明

頻度	表記
１２	オーム／真理／教
4	國林長／官／狙撃／事件
3	第／一／波
2	アウン／・／タン／・／スー／・／チー

（ｂ）候補単語辞書の説明

表記	品詞	詳細品詞
オーム真理教	名詞	普通名詞

（ｃ）登録前の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	普通名詞	オーム
名詞	普通名詞	真理
名詞	普通名詞	教
助詞	助詞「の」	の
名詞	人称名詞	信者
助詞	助詞「の」	の
名詞	人名	林
名詞	人名	春男
名詞	人称名詞	容疑者
助詞	格助詞	が
名詞	時詞	きょう
名詞	サ変名詞	逮捕
動詞語尾	サ変未然形	され
助動詞	れる	れ
助動詞	ます連用形	まし
助動詞	た終止連体形	た
その他	句点	。

【図１２】

未登録複合語頻度表を作成する場合の説明図（2）

（ａ）登録した後の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	普通名詞	オーム真理教
助詞	助詞「の」	の
名詞	人称名詞	信者
助詞	助詞「の」	の
名詞	人名	林
名詞	人名	泰男
名詞	人称名詞	容疑者
助詞	格助詞	が
名詞	時詞	きょう
名詞	サ変名詞	逮捕
動詞語尾	サ変未然形	され
助動詞	れる	
助動詞	ます連用形	まし
助動詞	た終止連体形	た
その他	句点	。

（ｂ）ユーザが修正した候補単語辞書の説明

表記	品詞	詳細品詞
オーム真理教	名詞	固有名詞

【図１４】

関連語を登録する場合の説明図（２）

（ａ）「國林長」を登録した場合の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	固有名詞	警察庁
助詞	助詞「の」	の
名詞	普通名詞	國林長
接尾語	接尾語	官
名詞	サ変名詞	狙撃
名詞	普通名詞	事件
助詞	助詞「の」	の
名詞	サ変名詞	捜査
助詞	格助詞	を
動詞	ラ行五段動詞	めぐ
動詞語尾	ラ行五段終止連	る
名詞	サ変名詞	対応
助詞	格助詞	が
形容動詞	ナ活用形容動詞	適切
形容動詞	形容動詞連用形	で
助動詞	ない	な
形容詞語	形容詞終止連体	い
その他	句点	。

（ｂ）「國林長官狙撃事件」を登録した場合の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	固有名詞	警察庁
助詞	助詞「の」	の
名詞	普通名詞	國林長官狙撃事件
助詞	助詞「の」	の
名詞	サ変名詞	捜査
助詞	格助詞	を
動詞	ラ行五段動詞	めぐ
動詞語尾	ラ行五段終止連	る
名詞	サ変名詞	対応
助詞	格助詞	が
形容動詞	ナ活用形容動詞	適切
形容動詞	形容動詞連用形	で
助動詞	ない	な
形容詞語	形容詞終止連体	い
その他	句点	。

【図１５】

関連語を登録する場合の説明図（３）

（ａ）ユーザが修正した候補単語辞書の説明

表記	品詞	詳細品詞
國林長官狙撃事件	名詞	固有名詞

（ｂ）登録前の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	未登録語	アウン
記号	記号	・
名詞	普通名詞	タン
記号	記号	・
名詞	未登録語	スー
記号	記号	・
名詞	未登録語	チー
接尾語	人称名詞接尾語	さん
助詞	助詞「の」	の
名詞	普通名詞	勢力
助詞	格助詞	と
助詞	提題助詞	は
名詞	普通名詞	一線
助詞	格助詞	を
動詞	サ変動詞	画し
動詞語尾	サ変連用形	てい
その他	た形接続語	る
動詞語尾	一段終止連体形	。
その他	句点	

【図１６】

関連語を登録する場合の説明図（４）

（ａ）「アウン・タン・スー・チー」を登録した場合の形態素解析結果の説明

品詞	詳細品詞	表記
名詞	普通名詞	アウン・タン・スー・チー
接尾語	人称名詞接尾語	さん
助詞	助詞「の」	の
名詞	普通名詞	勢力
助詞	格助詞	と
助詞	提題助詞	は
名詞	普通名詞	一線
助詞	格助詞	を
動詞	サ変動詞	画
動詞語尾	サ変連用形	し
その他	た形接続語	てい
動詞語尾	一段終止連体形	る
その他	句点	。

（ｂ）ユーザが修正した候補単語辞書の説明

表記	品詞	詳細品詞
アウン・タン・スー・チー	名詞	人名